DOCUMENT RESUME

ED 082 656 HE 004 682

AUTHOR Gillmore, Gerald M.
TITLE Estimates of Reliability Coefficients for Items and Subscales of the Illinois Course Evaluation Questionnaire. Report #341.
INSTITUTION Illinois Univ., Urbana. Office of Instructional Resources.
PUB DATE Aug 73
NOTE 42p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Attitudes; *Course Evaluation; *Evaluation; Evaluation Methods; *Higher Education; *Questionnaires; *Student Opinion
IDENTIFIERS *Illinois Course Evaluation Questionnaire

ABSTRACT

The major focus of this paper is on the reliability of the individual items, the subscales, and the total score of Form 66 of the Illinois Course Evaluation Questionnaire (CEQ). The CEQ is a Likert-type attitude questionnaire designed to elicit evaluative information from students about courses in which they are enrolled. Form 66 contains 50 items that are combined by unweighted averaging to form six subscales and a total score. "Stability" coefficients were estimated by three methods applied to six samples. All estimates indicated reasonably high reliabilities for classes. The magnitude of these reliabilities were discussed in the context of standard errors of measurement that was discussed, in turn, within the context of norming. (Author)

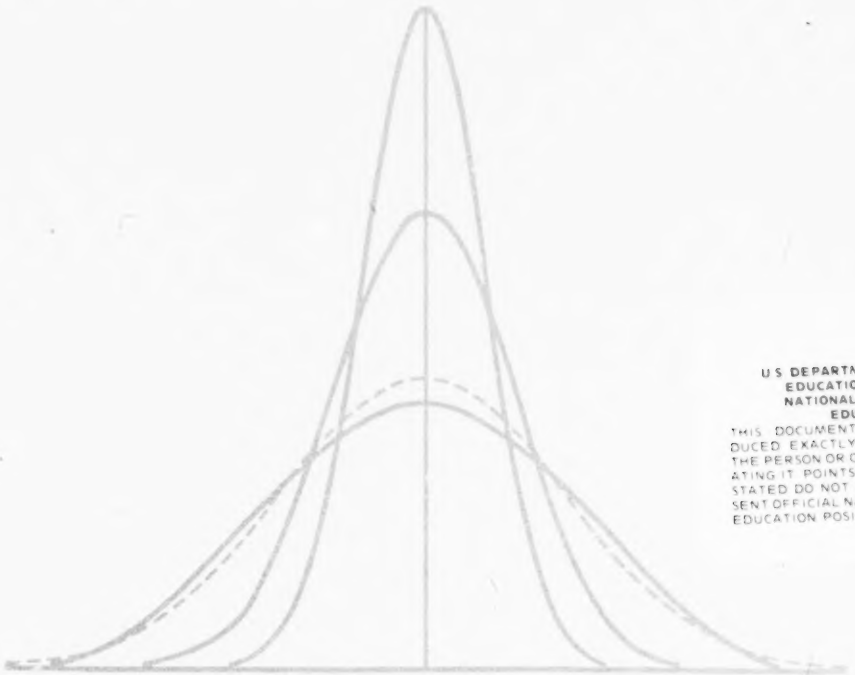
FILMED FROM BEST AVAILABLE COPY

Research Report

TITLE ESTIMATES OF RELIABILITY COEFFICIENTS FOR ITEMS AND SUBSCALES OF THE ILLINOIS COURSE EVALUATION QUESTIONNAIRE

AUTHOR(S) Gerald M. Gillmore

ED 082656



U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

REPORT # 341 DATE AUGUST, 1973

May be quoted in whole or in part if credit is given the source.

MEASUREMENT AND RESEARCH DIVISION - OFFICE OF INSTRUCTIONAL RESOURCES
307 ENGINEERING HALL - UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS 61801

HE004872

Abstract

The major focus of this paper is on the reliability of the individual items, the subscales, and the total score of Form 66 of the Illinois Course Evaluation Questionnaire (CEQ). The CEQ is a Likert-type attitude questionnaire designed to elicit evaluative information from students about courses in which they are enrolled. Form 66 contains 50 items, which are combined by unweighted averaging to form six subscales and a total score.

"Stability" coefficients were estimated by three methods applied to six samples. All estimates indicated reasonably high reliabilities for classes. The magnitude of these reliabilities were discussed in relation to number of students in a class.

"Equivalence" coefficients were presented for all subscales and discussed. Finally, reliabilities were discussed in the context of standard errors of measurement which was discussed, in turn, within the context of norming.

NOTE: The results of this paper can be used with CEQ Form 72 and Form 73 as it contains items and subscales found in Form 66.

ESTIMATES OF RELIABILITY COEFFICIENTS FOR ITEMS AND SUBSCALES OF THE ILLINOIS COURSE EVALUATION QUESTIONNAIRE

Gerald M. Gillmore

The major focus of this paper will be on the reliability of the individual items, the subscales, and the total score of Form 66 of the Illinois Course Evaluation Questionnaire (CEQ). Various methods of estimation will be discussed, and estimates using each method will be presented. Reliability estimates for the six subscales and for the total score will also be presented and discussed, as will estimates of the standard errors of measurement, especially as related to item and subscale norms.

The Illinois Course Evaluation Questionnaire (CEQ)

The CEQ is a Likert-type attitude questionnaire designed to elicit evaluative information from students about courses in which they are enrolled. Form 66 of the CEQ contains 50 items. The items, numbered as they appear on the form, are listed in Table 1. All items use the four response categories: Strongly Agree, Agree, Disagree, Strongly Disagree. Subsets of these items are combined by unweighted averaging to form six subscales. The names of the subscales and the list of items which form them are found in Table 2. A total score is also available by averaging over all 50 items. Details of the development of the instrument can be found in Spencer and Aleamoni (1969). The actual form can be found in Appendix A.

Reliability -- What Is It?

Reliability has been defined in a number of distinct but related ways. The thread of meaning throughout is the idea of "consistency" (Rozeboom, 1966, p. 375) or "repeatable" (Nunnally, 1967, p. 172). Reliability has a great deal of semantic overlap with the concept of "precision" as used

Table 1

The Items of the Illinois Course Evaluation Questionnaire

1.	I learn more when other teaching methods are used.
2.	It was a waste of time.
3.	Overall, the course was good.
4.	The textbook was very good.
5.	The instructor seemed to be interested in students as persons.
6.	More courses should be taught this way.
7.	The course held my interest.
8.	I would have preferred another method of teaching in this course.
9.	It was easy to remain attentive.
10.	The instructor did not synthesize, integrate or summarize effectively.
11.	Not much was gained by taking this course.
12.	The instructor encouraged the development of new viewpoints and appreciations.
13.	The course material seemed worthwhile.
14.	It was difficult to remain attentive.
15.	Instructor did not review promptly and in such a way that students could understand their weaknesses.
16.	Homework assignments were helpful in understanding the course.
17.	There was not enough student participation for this type of course.
18.	The instructor had a thorough knowledge of his subject matter.
19.	The content of the course was good.
20.	The course increased my general knowledge.
21.	The types of test questions used were good.
22.	Held my attention throughout the course.
23.	The demands of the students were not considered by the instructor.
24.	Uninteresting course.
25.	It was a very worthwhile course.
26.	Some things were not explained very well.
27.	The way in which this course was taught results in better student learning.
28.	The course material was too difficult.
29.	One of my poorest courses.
30.	Material in the course was easy to follow.
31.	The instructor seemed to consider teaching as a chore or routine activity.
32.	More outside reading is necessary.
33.	Course material was poorly organized.
34.	Course was not very helpful.
35.	It was quite interesting.
36.	I think that the course was taught quite well.
37.	I would prefer a different method of instruction.
38.	The pace of the course was too slow.
39.	At times I was confused.
40.	Excellent course content.
41.	The examinations were too difficult.
42.	Generally, the course was well organized.
43.	Ideas and concepts were developed too rapidly.
44.	The content of the course was too elementary.
45.	Some days I was not very interested in this course.
46.	It was quite boring.
47.	The instructor exhibited professional dignity and bearing in the classroom.
48.	Another method of instruction should have been employed.
49.	The course was quite useful.
50.	I would take another course that was taught this way.

Table 2

CEQ Items Grouped by Subscales

01. General Course Attitude (G.C.A.)

2. It was a waste of time.
3. Overall, the course was good.
11. Not much was gained by taking this course.
20. The course increased my general knowledge.
25. It was a very worthwhile course.
29. One of my poorest courses.
34. Course was not very helpful.
49. The course was quite useful.

02. Method of Instruction (M.I.)

1. I learn more when other teaching methods are used.
6. More courses should be taught this way.
8. I would have preferred another method of teaching in this course.
27. The way in which this course was taught results in better student learning.
36. I think that the course was taught quite well.
37. I would prefer a different method of instruction.
48. Another method of instruction should have been employed.
50. I would take another course that was taught this way.

03. Course Content (C.C.)

13. The course material seemed worthwhile.
19. The content of the course was good.
26. Some things were not explained very well.
28. The course material was too difficult.
30. Material in the course was easy to follow.
39. At times I was confused.
40. Excellent course content.
44. The content of the course was too elementary.

04. Interest and Attention (I.A.)

7. The course held my interest.
9. It was easy to remain attentive.
14. It was difficult to remain attentive.
22. Held my attention throughout the course.
24. Uninteresting course.
35. It was quite interesting.
45. Some days I was not very interested in this course.
46. It was quite boring.

05. Instructor (Instr.)

5. The instructor seemed to be interested in students as persons.
10. The instructor did not synthesize, integrate, or summarize effectively.
12. The instructor encouraged the development of new viewpoints and appreciations.
15. Instructor did not review promptly and in such a way that students could understand their weaknesses.
18. The instructor had a thorough knowledge of his subject matter.
23. The demands of the students were not considered by the instructor.
31. The instructor seemed to consider teaching as a chore or routine activity.
47. The instructor exhibited professional dignity and bearing in the classroom.

06. Specific Items (S.I.)

4. The textbook was very good.
16. Homework assignments were helpful in understanding the course.
17. There was not enough student participation for this type of course.
21. The types of test questions used were good.
32. More outside reading is necessary.
33. Course material was poorly organized.
38. The pace of the course was too slow.
41. The examinations were too difficult.
42. Generally, the course was well organized.
43. Ideas and concepts were developed too rapidly.

in the physical sciences (Gillmore and Stallings, 1971). Eisenhart (1968) defined precision as "... the typical closeness together of successive independent measurements of a single magnitude generated by repeated applications of the process under specified conditions" (p. 1201).

The social and behavioral sciences have a problem within this domain, however, which physical scientists typically do not share; namely, memory. A physical scientist can weigh the same block of lead repeatedly, for example, and obtain a collection or distribution of estimates of its weight. The precision of the measurements can be assessed in a reasonably straight-forward manner, essentially as a function of the "spread" of the distribution. However, if one wishes to assess the reliability of the measurement of an attitude held by an individual, he cannot repeatedly ask the same question with no intervening time. Human short-term memory is close enough to perfect to suspect little or no variation in responses, even though the measurement of the attitude may be potentially quite imprecise.

Two basic approaches are taken to circumvent this problem. First, multiple independent measurements can be taken simultaneously. This could be done by having multiple persons rating the same person on a given attribute or by asking the same person a series of somewhat different questions which all relate to the same attitude under investigation. Second, the same measurement can be taken twice, with an intervening time interval considered, on one hand, long enough to assure that memory of the first response is not a determiner of the second measurement and, on the other hand, short enough to assure that the thing being measured has not changed drastically in the interim.

There are also two basic types of reliability estimates. "A retest after an interval, using the identical test, indicates how stable scores are and, therefore, can be called a coeffieient of *stability*. The correlation

between two forms given virtually at the same time is a coefficient of *equivalence* showing how nearly two measures of the same trait agree " (Cronbach, 1951, p. 298).

The majority of this paper will be aimed at the former type of reliability. Stability is being emphasized for two reasons. The first reason is neither subtle nor profound. There is no way to assess the equivalency of a single item. Thus, we are left with the ability to look at equivalency only in the case of subscales. The second reason relates to purpose. In the context of the results of an instructor's course evaluation, the question to which estimates of stability relates is: To what extent are these results consistent with results which would have been obtained with the same course taught to an entirely different set of students from the same population? The question to which estimates of equivalency relates is: To what extent do the measurements of this aspect of teaching go together? To put the same question another way: To what extent do the measurements seem to be assessing a single underlying trait? For this paper, estimates of stability seem to be the more important of the two. However, data on the equivalence of the CEQ subscales will be presented and discussed subsequent to the presentation and discussion of the stability coefficients.

The Estimation of Reliability Coefficients

Generally, reliability coefficients must be estimated from data obtained from a sample of measurements or measurements from a sample of people. For purposes of the present paper, they must be estimated because the population of all instructors has not been measured. Indeed, no one can be sure what the population really is.

In the ideal situation, one tries to choose a sample completely randomly from a well-defined population.[1] Often, however, the ideal is not

---

[1]By randomly, we mean that the method of selection of a sample assures each member of the population an equal chance to become a member of the sample.

obtainable simply because of the impossibility or impracticality of selecting some members of the population for the sample.

The problem of a non-random sample is especially severe in the present study. Although Form 66 of the CEQ has had extensive use, especially at the University of Illinois but also at other institutions, its administration has never been mandatory. Most of the data used in this study was obtained from courses whose instructor chose to administer the CEQ.

The raters, i.e., students, who filled out the forms are not random either. Students do not enroll for courses by random selection. Furthermore, in some cases, the same students have undoubtedly rated several different instructors. In other cases, different instructors are rated by completely independent raters. In a few cases, the same students may have rated the same instructor more than once. These uncontrolled factors tend to contaminate the data in largely unknown ways.

The effect of these biases cannot be completely assessed and certainly not eliminated. However, different methods of estimation can minimize some of the biases. Using multiple methods also can give some confidence that the range of reliability coefficients captures the "true" reliability. Finally, if the varying methods and samples give similar results, more confidence can be given to the accuracy of those results. It is for these reasons that the multiple method-multiple sample approach was adopted. In all, six different stability estimates will be presented for each item and five for each subscale and the total score. Three equivalence coefficients for each subscale will also be presented following the presentation of the stability coefficients.

Stability Coefficients for the Items and Subscales

Method 1: The Intraclass Correlation Coefficient

One can look at the set of students who rate each instructor as raters. Each section which is rated can be looked at as a group. Then, within analysis

of variance language, the situation with many sections being rated is basically a one-way design with students nested within groups. Since it makes little sense to generalize all results to the particular set of sections in which the CEQ has been used, sections should probably be considered as a random sample of all possible sections and thereby adopt a random effects or variance components model. (Computationally, in a one-way design, the two designs are identical).

Using this model, one can partition the total sum of squares from actual data into that due to groups (sections) and that due to raters within groups. The reliability of the raters can be estimated from this data by use of the intraclass correlation. (For derivations of the intraclass correlation, see Ebel, 1951). In this case, since differences in level of rating between raters does make a difference in the evaluation an instructor receives, the "between raters" variance should be part of the error term. "But, if decisions are made in practice by comparing single 'raw' scores assigned to different pupils (instructors) by different raters, or *by comparing averages which come from different groups of raters,* then the 'between-raters' variance should be included as part of the error term" (Italics mine) (Ebel, 1951, p. 412). Since we are assuming completely independent raters for each section, the between raters variance is, indeed, automatically a part of the error term.

If we were interested in the reliability of the rating of an individual "average" rater, the formula for the intraclass correlation of relevance would be as follows:

$$r = \frac{MSB - MSW}{MSB + (K-1)(MSW)} \qquad (1)$$

where MSB refers to the Mean Square between groups or sections,

MSW refers to the Mean Square between raters within groups,

and K is the number of raters per group.

When K is not constant, it can be estimated with the following formula:

$$K = \frac{1}{n - 1} \left(\Sigma k_i - \frac{\Sigma k_i^2}{\Sigma k_i}\right)$$

where n = number of groups and k_i is the number of raters in the i^{th} group. However, we are interested in the reliability of the ratings of the total set of raters for each section, since instructors are evaluated in terms of class means rather than individual ratings. Thus, the appropriate formula becomes as follows:

$$r' = \frac{MSB - MSW}{MSB} \qquad (2)$$

r' is very close to the value of r inflated by the Spearman-Brown Prophecy Formula with the n, which usually refers to the increased length of a test, in this case referring to the average number of raters. The Spearman-Brown Prophecy Formula is as follows:

$$r' = \frac{nr}{(n - 1)\, r + 1} \qquad (3)$$

The intraclass correlation was computed on three different samples of CEQ results.[2]

Sample 1

The most alluring aspect of Sample 1 is its immensity. The sample contains data from the 5,346 sections whose instructors gave the CEQ between the years 1966 and 1970. Of these sections, 2,782 were taught at the University

---

[2]The *F* value from the analysis of variance can be computed from the intraclass correlation as follows: $F = \frac{1}{1 - r'}$. One can note that if the reliability is zero, F = 1, its expected value. One can also note that if the reliability is perfect, F is infinite.

of Illinois (Urbana campus). The remaining sections were taught at 18 different colleges and universities across the country. In all, these data are ratings from 105,576 raters (not necessarily all different).

One assumption of the model which has been adopted is that the sample of sections is randomly chosen from the entire population. This is, of course, not true of these data since they were collected from volunteers. Furthermore, some instructors and courses are represented more than once. This could be possible with random selection, however, with the non-random method used for this sample, it probably worsens any bias there might be.

Also, according to the model, raters should be randomly assigned to sections. Furthermore, no rater should rate more than one course since otherwise dependencies will be evident in the data, i.e., the correlation built in between a rater's rating of one course and another. These data do not strictly satisfy either requirement. It is not always clear why students choose the courses they do, but it is clearly not random. Furthermore, there is little doubt that some students rated more than one course within the sample.

Given these obvious limitations of the data, even the most imperceptive of readers might reasonably inquire as to why analysis was carried out at all. Beyond the natural passion of statisticians for large amounts of data, the biases may not be as destructive as one might at first suspect. First, the greatest effect of the volunteer nature of the sample would probably be to restrict the range of responses. For example, one might expect really poor teachers to tend not to give the instrument, and thus, one tail of the distribution would be smaller in the sample than in the population. It is difficult for this researcher to conceive of a reason to expect the bias of the sample to increase the range unless one suspects that good and bad instructors tend to give the instrument, while mediocre instructors tend not to. However, the distribution of the data so closely approximates the

normal distribution (see Gillmore, 1971), this does not look likely. The result of restriction of range is characteristically to lower reliabilities (more basically, correlation). In the present context, the restriction of range clearly lowers the mean square between groups, but probably does not effect the within group mean square. Thus, the effect of this bias is probably to lower the reliability estimates rather than to inflate them.

Similarily, the effect of having an instructor or course rated more than once, but treating the data as if it were independent ratings, would seem to have a depressing effect on the reliability estimates. If two independent sets of raters rate the same instructor or course, the resulting ratings would certainly tend to be closer together than if two independent sets of raters rated two different instructors or courses. This again would suggest a reduction of the between groups sum of squares without an accompanying reduction of within group sum of squares.

Finally, the effect of having some of the same students rating two different courses or instructors would also seem not to have much effect. Consider the case where every student rates every instructor or course. This is a completely crossed two-way analysis of variance design with students as a random factor and instructors either random or fixed. In this case, the total sum of squares is partitioned into that due to instructors, to students, and to the instructor by student interaction. The typical error term for instructors is the latter. However, for the appropriate intraclass coefficient, the between raters variance is also part of the error. Thus, the two analyses would seem to be equivalent.

The last bias mentioned above was the lack of random assignment of raters to instructors. It might be pointed out that random assignment would make no sense in the educational setting in which we are working, but that is only saying that the model does not fit. Unfortunately, the effect of this lack of fit is not clear, nor can this researcher make any very intelligent guesses.

All of these influences in combination would seem to give some confidence that empirically determined reliability estimates by the method described above would not tend to be overestimates; indeed they might be more aptly considered underestimates.

The resulting stability coefficients for Sample 1 are found in Table 3. Because of the large amount of data, a completely accurate computation of the mean square between and mean square within for Formula 2 was not feasible. The average section size was used to compute the mean squares rather than using the correct weighted average. Also, the stability estimates for subscales and the total score were not calculated for this sample. They were calculated for all other samples. One can note that the coefficients for items range from .756 for Item 32 to .911 for Item 4. The mean of the 50 reliability estimates is .854.

Sample 2 and 3

One of the problems inherent in Sample 1 was that instructors could appear more than once in the sample. Similarly, courses could also appear more than once. To alleviate this problem, a sample was randomly chosen from University of Illinois (Urbana campus) courses taught fall term, 1971-72, who used the CEQ. The sample contained 200 courses, however, no course or instructor was allowed to appear in the sample more than once. A second sample of identical size was chosen for purposes of replication. The only additional criterion for exclusion was that no particular section could appear in the second sample which had appeared in the first. Thus, Sample 2 and 3 were independent random samples within the limits mentioned above. (The non-representativeness of the population from which the sample was drawn still remains a shortcoming.) The average number of students per class was 34.92 for Sample 1 and 28.15 for Sample 2.

Table 3

Reliability (Stability) Estimates for the Items, Subscales, and Total Scores of the CEQ by Three Different Methods and Six Different Samples.*

Item	Method 1 - Intraclass			Method 2 Test-Retest	Method 3 - Split-half	
	Sample 1 (N=5346)	Sample 2 (N=200)	Sample 3 (N=200)	Sample 4 (N=103 Pairs)	Sample 5 (N=103)	Sample 6 (N=103)
1	834	868	849	652	880	806
2	846	879	889	595	853	828
3	864	911	905	631	850	877
4	911	932	899	711	911	851
5	883	910	910	817	871	878
6	881	927	904	658	895	847
7	870	907	905	693	859	856
8	853	892	874	682	837	789
9	880	927	909	669	901	873
10	835	908	878	688	839	842
11	843	870	887	636	844	824
12	864	913	907	701	754	813
13	837	874	863	614	749	789
14	874	920	906	662	869	843
15	827	892	876	589	799	812
16	877	882	866	720	793	862
17	874	898	852	701	721	791
18	852	913	888	747	828	839
19	842	888	885	599	775	849
20	819	852	845	690	797	733
21	885	884	876	621	850	824
22	876	914	902	677	885	871
23	828	877	871	670	845	747
24	860	901	908	661	850	858
25	869	895	904	654	819	867
26	864	924	887	736	862	905
27	877	922	896	691	867	845
28	831	872	852	582	731	730
29	834	896	897	652	771	830
30	864	909	890	711	812	858
31	846	888	892	654	750	863
32	756	845	731	533	686	669
33	839	889	856	717	817	849
34	845	875	874	660	817	832
35	868	917	914	696	863	905
36	886	924	915	758	881	88
37	855	896	865	738	830	830
38	804	815	808	725	801	828
39	885	924	890	783	889	885
40	867	916	903	699	825	830
41	908	901	895	725	847	824
42	843	887	870	639	821	833
43	839	901	866	689	781	841
44	810	795	731	548	543	686
45	855	888	850	576	826	757

Table 3 (cont.)

Item	Method 1 - Intraclass			Method 2 Test-Retest	Method 3 - Split-half	
	Sample 1 (N=5346)	Sample 2 (N=200)	Sample 3 (N=200)	Sample 4 (N=103 Pairs)	Sample 5 (N=103)	Sample 6 (N=103)
46	855	895	899	685	824	854
47	817	854	812	576	730	751
48	854	896	869	713	829	838
49	859	889	885	669	817	839
50	869	905	897	671	861	834
Subscales						
G.C.A.		916	918	698	875	893
M.I.		931	914	733	899	871
C.C.		940	920	725	875	886
I.A.		933	926	704	900	893
Instr.		938	931	730	864	883
S.I.		923	895	697	867	903
Total		945	932	728	900	906

*Decimal points have been eliminated for ease of reading in this table and all subsequent tables.

The intraclass reliability coefficients calculated on these two samples appear in Table 3. The range of estimates for the items for Sample 2 was from .795 for item 44 to .905 for item 50. The range for the items for Sample 3 was from .731 for item 32 and 44 to .897 for item 50. The average stability coefficient for Sample 2 was .893, the average for Sample 3 was .876.

Also found in Table 3 are the results for the subscales and the total scores. In this case and subsequently, the stability coefficients are computed on the subscale means and the means of the total scale.

Method 2: Test-Retest

The second method of reliability estimation was similar to the test-retest method. Usually, in test-retest, the instrument is administered twice to the same group of subjects, with a period of time intervening judged to be long enough that subjects have forgotten specific responses they had made but short enough that true changes in the variable being measured would not be expected.

Since the same set of students do not take the same course by the same instructor twice (except possibly the very small subset who fail the first time), the traditional test-retest method was not possible to implement. A variant of this method was possible however.

Frequently, the same instructor may teach two or more sections of the same course. This procedure is most common in lower level courses. If the same instructor teaches two sections of the same class, and uses the CEQ in both, one would expect that the two sets of ratings would be much more similar than, say, two different instructors teaching two different courses; that is, if the instrument is reliable. If, on the other hand, the ratings of the same instructor teaching two sections of the same course were not similar, the reliability of the instrument would be highly questionable. The correlation of the means of the two sets of ratings, correlated over instructors who taught two sections of the same course, can be considered a reliability estimate.

There is a problem with this estimate of reliability, however. As with twin studies, which mean of an instructor goes into which of the two groups, is not clear. In the traditional test-retest method, the outcome of the first testing is correlated with the outcome of the second testing. However, it makes little sense to correlate the means of the first section that meets in a day with the means of the second section that meets in a day. On the other hand, random placement of sections into groups results in a variability of resulting correlations which need not be present. In the worst possible case, a researcher could do much to influence the size of a correlation by post hoc arrangement.

The proper way to alleviate this problem is to create a symmetric table (Treloar, 1942, pp. 11-13). For this method, each pair of observations is entered into the data matrix twice, once in each order. The correlations which result from this method are reliability coefficients (Jensen, 1971). They are a form of the intraclass correlation. However, they should be considered as "lower-bounds" to a test-retest reliability. We can be reasonably confident that the "true" test-retest correlation is not lower than these coefficients. There are two reasons for this statement. First, correlations computed from a symmetric matrix are smaller than correlations resulting from *any* combination of the same values in a non-symmetric matrix. Second, and probably more important, the comparisons are made between pairs of sections which differ in many important ways. The most important is different sets of raters. Other differences are time of day, size of the class, fatigue or practice effects on the instructor, etc. Thus, these stability coefficients are expected to be smaller than those presented previously.

Sample 4

During the academic year, 1970-71, a large proportion of instructors at the University of North Carolina, Greensburo (U.N.C.), used the CEQ. As a by-product, many instructors who taught two or more sections of the same course

used the CEQ for these sections. In all, 103 different instructors (and 103 different courses) fell within this category. When an instructor taught more than two sections of the same course, two of the sections were randomly chosen. Thus, the reliability was calculated, using a symmetric matrix for 103 instructors and courses, with two sections of each.

The results of application of Method 2 on Sample 4 can be found in Table 3 for all items. The range of reliabilities are from .533 for Item 32 to .817 for Item 5. The average reliability over the 50 items was .671. The results for the six subscales are also found in Table 3 as is the result for the total instrument.

Method 3: Split-half

The split-half method of calculating reliability is most commonly applied to situations in which a group of subjects respond to an instrument with many items. These items are typically attempting to measure a construct, such as knowledge in an achievement test, or a specific attitude in an attitude questionnaire. The items are split into two groups, usually the odd numbered items go into one group, the even numbered items go into the other. The means of the two groups are then correlated over subjects. Finally,the resulting correlation is raised by use of the Spearman-Brown Prophecy Formula for tests of double length[3], which is then a measure of equivalency of the set of items.

---

[3]The Spearman-Brown Prophecy Formula for tests of double length is as follows:

$$r_{tt} = \frac{2r_{ii}}{1 + r_{ii}}$$

where r_{ii} is the correlation between the two halves, and r_{tt} is the total reliability.

This same procedure can be followed for a single item with multiple raters. Raters can be split into two groups, and the average of the two groups correlated over sections. This result also needs to be raised by the Spearman-Brown Formula to become a reliability estimate. However, the result of applying the split-half method to raters is a *stability* coefficient rather than an equivalence coefficient.

Sample 5 and 6

The 103 pairs of sections used in Sample 4 also comprised Samples 5 and 6. One member of each pair was arbitrarily assigned to Sample 5, the other to Sample 6. Then within the two samples, the raters from each section were divided into two groups by means of an odd-even split based on the way the data were naturally ordered. Then, the means of the odd raters were correlated with the means of the even raters for each item within each course. Finally, the resulting correlations were corrected by means of the Spearman-Brown Prophecy Formula. The results are found in Table 3. For Sample 5, the stability coefficients for items ranged from .543 for Item 44 to .911 for Item 4. For Sample 6, the stability coefficient for items ranged from .669 for Item 32 to .905 for Items 26 and 35. The averages were .819 and .829 for Sample 5 and Sample 6, respectively.

Consistency of Stability Estimates

At this point, two considerations seem warranted. First, is there any consistency among the reliability estimates? Essentially, we are asking a seldom asked question: Are the stability coefficients reliable? If they are not, we would be hard-pressed to justify faith in their veracity. Secondly, given an adequate degree of consistency, what do they indicate? Or, to put this question into an over-simplified form: Are the items of the CEQ reliable?

To address the reliability of the reliabilities question, the six different reliability estimates were intercorrelated over the fifty items. (The reader should note that this is not a recommended procedure and the results should be interpreted with some caution.) The results are presented in Table 4. Subscales were not included in this analysis because of the small amount of variation in their stability estimates within methods. As can be seen, the correlations among the various techniques are reasonably large. The "test-retest" method tended to show the least agreement with the other methods, but even these correlations were .499 and above.

As an overall assessment of the consistency of the reliability estimates, the average off-diagonal correlation was calculated (.672), and corrected by the Spearman-Brown Prophecy Formula using a test of length six times the original. The resulting value, which is an estimate of the alpha coefficient, was .925.[4] We take this to suggest a consistency high enough to allow faith in our results. However, there still may be a source of systematic error which could affect all methods similarly and, therefore, not show up in the intercorrelations. An example of one such source is the volunteer nature of all the data.

In evaluating the magnitude of the reported reliability estimates, it must be noted that there are degrees of reliability, and while zero reliability, (i.e., complete unreliability) makes measurements unsuitable for any use, varying degrees of reliability are adequate for different situations, depending essentially on the fineness of discrimination needed, the supplemental data available, and the importance of resulting decisions. Thus, in reality, one cannot evaluate the stability coefficients in any abstract sense. However, in considering the practical meaning of the various indices of reliability

---

[4]The alpha coefficient is a measure of equivalence which will be discussed in the subsequent section.

Table 4

The Correlations Among Reliability Estimates

Samples	Method 1			Method 2	Method 3	
	1	2	3	4	5	6
1	1000					
2	765	1000				
3	765	840	1000			
4	555	517	499	1000		
5	632	698	751	525	1000	
6	653	697	807	665	710	1000

which have been described, another almost paradoxical issue comes to the fore; that of class size.

Typically in studies of reliability, numbers of subjects have little effect. This is because, in general, the magnitude of a correlation coefficient is unaffected by sample size other than the fact that there is more variability associated with smaller sample sizes. However, the length of the measuring device has an effect. Generally, reliability increases with increased length, other things being equal. However, since in the present situation we use raters analogously to items, class size does make a difference in the reliability of the class mean on a particular item. The reliability of a mean rating on an item in a class of 100 will generally be greater than for a class of 10, though not ten times greater. So what has been presented above are stability coefficients for average size classes! To get a fix on what effect the class size variable has, we can go back to the intraclass reliability coefficient for an individual rater. We present these values as calculated from Sample 1, 2, and 3 in Table 5. In all three samples, the smallest intraclass value, rounding off to two places, is .09 for Items 32 and 44 in Sample 3. The largest value is .33 for Item 41 in Sample 1. The average value is about .21.

These values essentially represent what the reliability of an individual rater would be. To continue our analogy with items, these reliabilities are comparable to the reliability of a single item, or the average off-diagonal correlation among items. And, just as a reliability for a set of items can be computed by applying the Spearman-Brown Prophecy Formula to average off-diagonal interitem correlations and the number of items[5], so can the reliability of

---

[5] $\text{Reliability} = \frac{n\bar{r}}{(n-1)\bar{r}+1}$

where n = number of items and
$\bar{r}$ = average off-diagonal correlation.

Table 5

Intraclass Correlations for Individual Raters

Computed on Sample 1, 2, and 3

Item	Sample 1	Sample 2	Sample 3
1	202	155	161
2	217	168	216
3	243	222	246
4	342	276	232
5	276	219	258
6	272	262	244
7	253	214	246
8	227	187	192
9	270	260	255
10	205	215	197
11	214	157	211
12	244	226	250
13	206	162	178
14	259	242	249
15	194	188	194
16	265	173	182
17	259	197	164
18	226	226	215
19	212	181	209
20	186	138	158
21	280	175	195
22	264	229	241
23	196	165	188
24	202	252	237
25	192	244	252
26	254	213	244
27	248	229	266
28	160	165	199
29	194	230	202
30	218	217	244
31	181	221	217
32	132	085	136
33	183	170	208
34	163	192	216
35	235	268	250
36	254	269	282
37	194	180	231
38	109	126	172
39	252	217	280
40	234	243	247

Table 5 (cont.)

Item	Sample 1	Sample 2	Sample 3
41	202	227	334
42	179	186	214
43	203	182	209
44	097	085	178
45	181	163	229
46	193	235	231
47	140	129	184
48	194	186	229
49	182	209	235
50	209	231	251
Subscales			
G.C.A.	--	233	279
M.I.	--	272	268
C.C.	--	303	282
I.A.	--	281	301
Instr.	--	296	317
S.I.	--	250	226
Total	--	322	320

ratings be assessed by applying the Spearman-Brown Prophecy to the intraclass correlation for an individual rater and the total number of raters. By this means, we can get an idea of what the reliability for a given item will be within a class of a certain size.

To get an idea of the effect of class size, the results of application of the Spearman-Brown Prophecy Formula plotted as a function of classes of size 1-40, for the high, low, and average intraclass correlation mentioned above is in Figure 1. As can be seen, a reliability of .80 is achieved for even the lowest intraclass correlation with a class of 40 or above. For the average value, the .80 magnitude is reached by class size of 15. Finally, for the highest intraclass correlation, a class size of 11 is sufficient to reach .80.

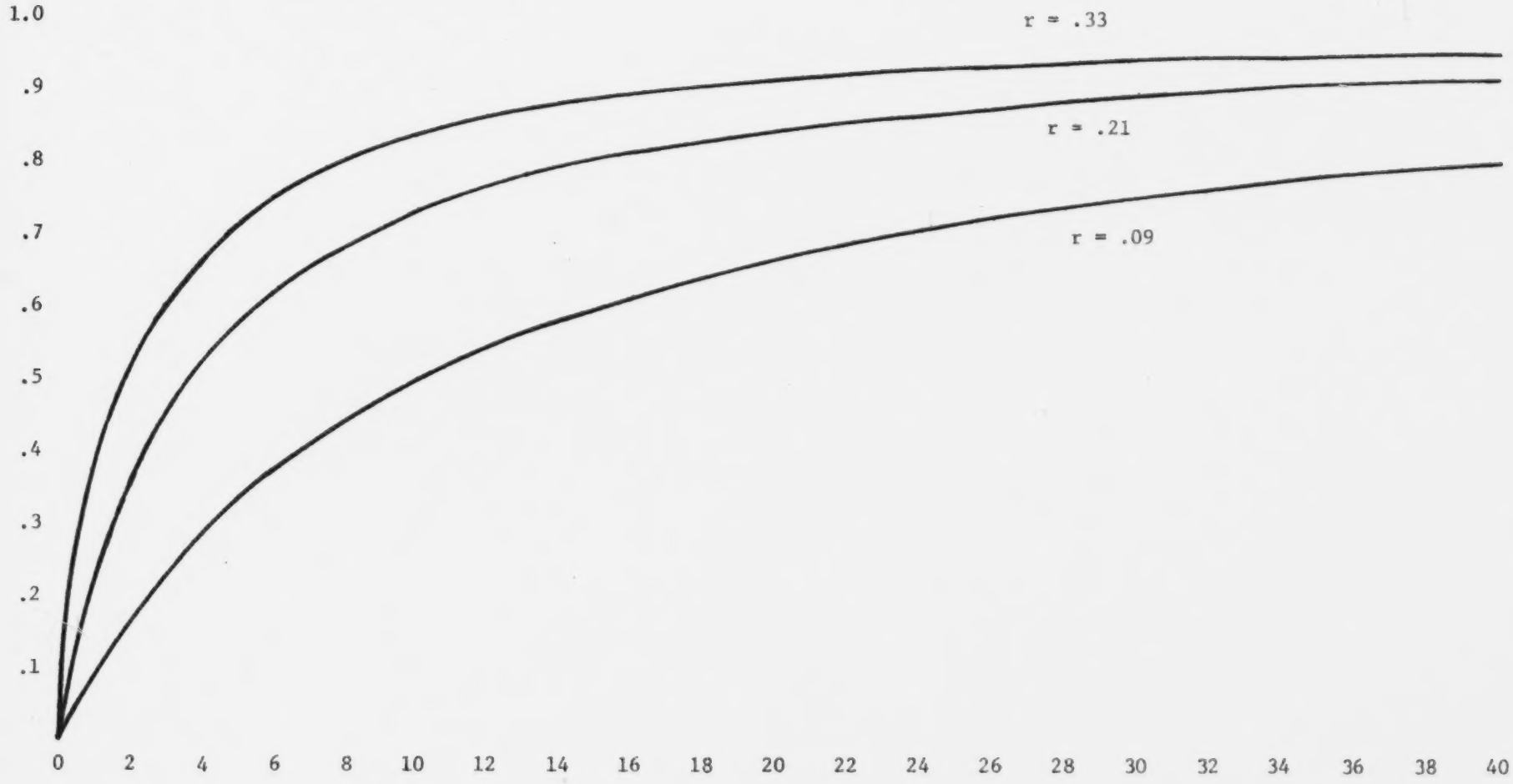
In most contexts, a stability coefficient for a single item of .80 would seem to be sufficient. Indeed, .70 may very well be an acceptable figure. If so, it is obtained for all but the smallest class for all but the lowest intraclass correlations. In general, it seems fair to say that the items of the CEQ have adequate stability, although one should realize that as class size decreases, interpretations may become more tenuous.

The same general statement can be made for the subscales. However, the stability coefficients for subscales definitely tend to be as high or higher than any of those for the items, as would be expected. Thus, somewhat more confidence can be lent to the results of subscales than items.

Equivalence Coefficients for the Subscales

Subscales by their very nature are made up of collections of items. Subscale scores, or subscores, are calculated by summing or averaging the scores of the items contained in the subscore. Thus, it is reasonable to question the relatedness, or equivalence, of the member items. For example, if they all measure completely different attributes, the single number representing the subscore has little meaning.

Figure 1. Reliability as a Function of Class Size for Three Intraclass Correlation Coefficients for Individual Raters



Coefficient alpha is a reliability estimate assessing the equivalence of a set of items (Cronbach, 1951). The formula for alpha is as follows

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\Sigma V_i}{V_t}\right)$$

where k is the number of items,

V_i is the variance of the i^{th} item, and

V_t is the variance of the sum of the items.

Cronbach (1951) has shown that coefficient alpha is the average of all possible split-half reliability coefficients. Furthermore, in the case where all items have equal variances, coefficient alpha is exactly equal to computing the average off-diagonal correlation among items and entering it into the Spearman-Brown Prophecy Formula (see footnote 5), where n is the number of items.

Coefficient alpha was computed for all subscales for Samples 1, 2, and 3. The results are found in Table 6. Since the unit of analysis is sections, in all cases section means rather than individual student ratings were entered into this analysis.

The equivalence coefficients for all three samples are very consistent within subscales. Three subscales, General Course Attitude, Method of Instruction, and Interest and Attention, show very high reliability. The reliabilities of Course Content and Instructor are moderately high. Finally, Specific Items has a considerably lower reliability than the others.

If one carefully reads the items which form each subscale (Table 2), he can see reasons for the existence of these differences. The three subscales which show extreme reliabilities all contain nearly equivalent items. Differences among the items are mainly subtle wording changes, i.e. synonyms. The Instructor subscale, on the other hand, contains somewhat dissimilar questions,

Table 6

Correlations Among the Items of the
General Course Attitude Subscale*

Items	2	3	11	20	25	29	34
Sample 1							
3	882						
11	905	883					
20	783	770	819				
25	871	895	911	831			
29	752	786	770	706	780		
34	869	867	907	831	914	786	
49	850	863	897	830	931	777	919
Sample 2							
3	926						
11	948	938					
20	863	879	892				
25	922	928	945	902			
29	914	941	923	854	919		
34	928	928	950	896	951	934	
49	907	904	935	891	952	905	952
Sample 3							
3	931						
11	926	876					
20	864	843	869				
25	925	925	921	884			
29	879	911	853	806	882		
34	905	887	928	892	929	866	
49	910	899	917	898	961	864	938

*Items 2, 11, 29, and 34 have been reverse scored such that a high score for all items represents a favorable response.

although all relating to the instructor of the course. The Course Content subscale has clusters of meaning. Items 13, 19, and 40 are very similar, all dealing with the value of course material. Items 28, 30, 39, and 44 all deal with the relative difficulty of the course material. Finally, Item 26, and to some extent 39, deals with the instructor's explanation of the course material. Thus, although all items relate the course content, they are not all homogeneous in content, hence, the lower reliability.

The Specific Items subscale is an even more extreme case of non-homogeneous item content. A whole collection of course related topics is included in the items of the subscale and, therefore, less equivalence is the result. Indeed, this collection of items stretches the definition of a subscale.

To get an idea of the relationship among items, the correlations among items within subscales are presented in Tables 7 through 12. Correlation matrices from all three samples are included. Again, these correlations are computed over section means. The preceding comments concerning the content of the items are generally borne out by the structure of these matrices. One can note consistently high correlations among the items of the General Course Attitude, Method of Instruction, and Interest and Attention subscales. Correlations among the Instructor items are consistent but lower. The Course Content subscales show definite clustering as suggested above. Finally, the Specific Items subscale contains generally low correlations, but some clustering can be seen.

Conclusion - The Standard Error of Measurement

This report will conclude both a brief discussion of the meaning of the reliability estimates which have been presented in this paper in terms of the Standard Error of Measurement (S.E.M.).

S.E.M. is essentially the standard deviation of a hypothetical distribution of observed scores around a "true" score; essentially, a distribution of

Table 7

Correlations Among the Items of the Method of Instruction Subscale*

Items	1	6	8	27	36	37	48
Sample 1							
6	832						
8	892	887					
27	848	915	894				
36	819	860	854	871			
37	882	885	925	900	883		
48	885	895	931	910	879	942	
50	802	909	851	883	835	859	877
Sample 2							
6	899						
8	911	934					
27	870	945	924				
36	867	912	896	909			
37	902	931	958	933	898		
48	911	930	947	915	899	946	
50	883	942	923	938	902	922	928
Sample 3							
6	901						
8	925	939					
27	887	946	919				
36	901	909	901	902			
37	931	936	956	927	921		
48	928	917	946	905	908	951	
50	887	935	905	932	899	907	895

*Items 1, 8, 37, and 48 have been reverse scored such that a high score for all items represents a favorable response.

Table 8

Correlations Among the Items of the

Course Content Subscale*

Items	13	19	26	28	30	39	40
Sample 1							
19	893						
26	514	588					
28	446	444	525				
30	452	484	631	800			
39	327	366	644	711	809		
40	873	900	627	448	506	394	
44	438	459	190	-188	-210	-248	441
Sample 2							
19	916						
26	595	682					
28	473	495	605				
30	520	570	767	780			
39	467	512	763	755	864		
40	914	943	713	507	615	546	
44	474	490	195	-108	-122	-094	468
Sample 3							
19	924						
26	603	694					
28	611	612	673				
30	582	621	784	836			
39	508	557	778	806	904		
40	896	928	711	601	638	573	
44	371	393	229	-050	-096	-138	415

*Items 26, 28, 39, and 44 have been reverse scored such that a high score for all items represents a favorable response.

Table 9

Correlations Among the Items of the Interest and Attention Subscale*

Items	7	9	14	22	24	35	45
Sample 1							
9	908						
14	908	959					
22	929	941	939				
24	917	842	859	878			
35	932	869	876	904	931		
45	853	852	859	875	816	841	
46	915	883	895	901	920	910	841
Sample 2							
9	950						
14	940	974					
22	959	964	960				
24	953	918	919	923			
35	970	923	911	929	950		
45	879	900	896	904	844	867	
46	963	944	950	943	943	948	885
Sample 3							
9	930						
14	918	970					
22	944	960	945				
24	942	893	889	905			
35	956	918	908	943	949		
45	876	887	877	916	849	877	
46	937	923	921	938	939	939	882

*Items 14, 24, 45, and 46 have been reverse scored such that a high score for all items represents a favorable response.

Table 10

Correlations Among the Items of the Instructor Subscale*

Items	5	10	12	15	18	23	31
Sample 1							
10	570						
12	706	546					
15	645	788	556				
18	411	589	458	462			
23	771	632	696	700	445		
31	760	668	685	641	596	713	
47	485	540	478	457	607	491	633
Sample 2							
10	639						
12	725	591					
15	678	833	651				
18	448	626	459	531			
23	770	681	807	752	478		
31	804	705	749	745	611	769	
47	471	627	476	531	693	485	629
Sample 3							
10	605						
12	670	557					
15	705	845	506				
18	423	647	406	545			
23	791	621	706	707	420		
31	797	733	756	698	630	797	
47	418	586	425	437	707	411	596

*Items 10, 15, 23, and 31 have been reverse scored such that a high score for all items represents a favorable response.

Table 11

Correlations Among the Items of the Specific Items Subscale*

Items	4	16	17	21	32	33	38	41	42
Sample 1									
16	351								
17	102	202							
21	313	430	303						
32	153	085	187	088					
33	347	371	378	491	287				
38	173	306	195	214	279	393			
41	165	231	321	594	044	339	-069		
42	330	398	400	509	257	910	417	329	
43	200	144	458	349	162	469	-198	539	457
Sample 2									
16	351								
17	229	322							
21	411	473	404						
32	299	158	042	204					
33	440	407	453	549	356				
38	246	319	277	310	322	423			
41	282	253	498	677	132	445	111		
42	441	398	508	564	287	935	416	451	
43	356	109	496	417	248	489	-079	617	491
Sample 3									
16	279								
17	250	376							
21	398	500	509						
32	341	131	184	206					
33	514	474	445	633	346				
38	240	325	251	222	152	501			
41	308	370	547	764	174	455	-050		
42	523	497	461	646	339	937	474	468	
43	366	279	537	631	297	518	-152	793	516

*Items 17, 32, 33, 38, 41, and 43 have been reverse scored such that a high score for all items represents a favorable response.

Table 12

Standardized Cutoff Scores for Decile Norms

Decile	Standardized Cutoff Score
0	
	-1.28
1	
	- .84
2	
	- .525
3	
	- .255
4	
	0
5	
	+ .255
6	
	+ .525
7	
	+ .84
8	
	+1.28
9	

measurement errors. A "true" score is an abstract quantity defined in various ways but essentially indicating what the score of an entity on a given variable *really is*. True scores are, of course, not directly measureable. If they were, there would be no need to estimate reliability coefficients.

The S.E.M. is important in that it gives some notions as to how close observed scores are likely to be to the "true" score.[6] For example, if one is making differential decisions on the basis of individual scores, one would hope for a small S.E.M. As the S.E.M. increases, his decisions are more apt to be due to measurement error than real differences.

The formula for S.E.M. is as follows:

$$S.E.M. = S\sqrt{1 - r} \qquad (4)$$

where S is the standard deviation of the observed scores and

r is the reliability.

The theoretical distribution of observed scores tends to be normal, with a mean at the true score and a standard deviation equal to the S.E.M. Consequently, about 68 percent of the observed scores will be within one S.E.M. of the true score in either direction. About 95 percent of the observed scores will be within two S.E.M.'s in either direction.

All of the S.E.M.'s for the various reliability estimates for items and subscales will not be presented. Rather some general notions will be suggested in the context of norming, since S.E.M.'s are probably most important when considering comparative judgments.

Currently, the CEQ norms are presented in the form of deciles. Instructors can get a decile rank of zero thru nine, with zero designating the lowest ten percent of the distribution of previous CEQ users, one designating from the

---

[6]It is important to note that an S.E.M. does not indicate how close a specific observed score is to the true score, since that observed score could lie anywhere in the distribution.

tenth to twentieth percentiles, etc. The CEQ deciles are computed by use of normal approximations (Gillmore, 1972).

Decile cutoff scores can be standardized, i.e., converted to z scores. Furthermore, since standard scores have a mean of zero and a standard deviation of one, the formula for S.E.M. (Formula 4) for standardized variables becomes:

$$S.E.M. = \sqrt{1 - r}$$

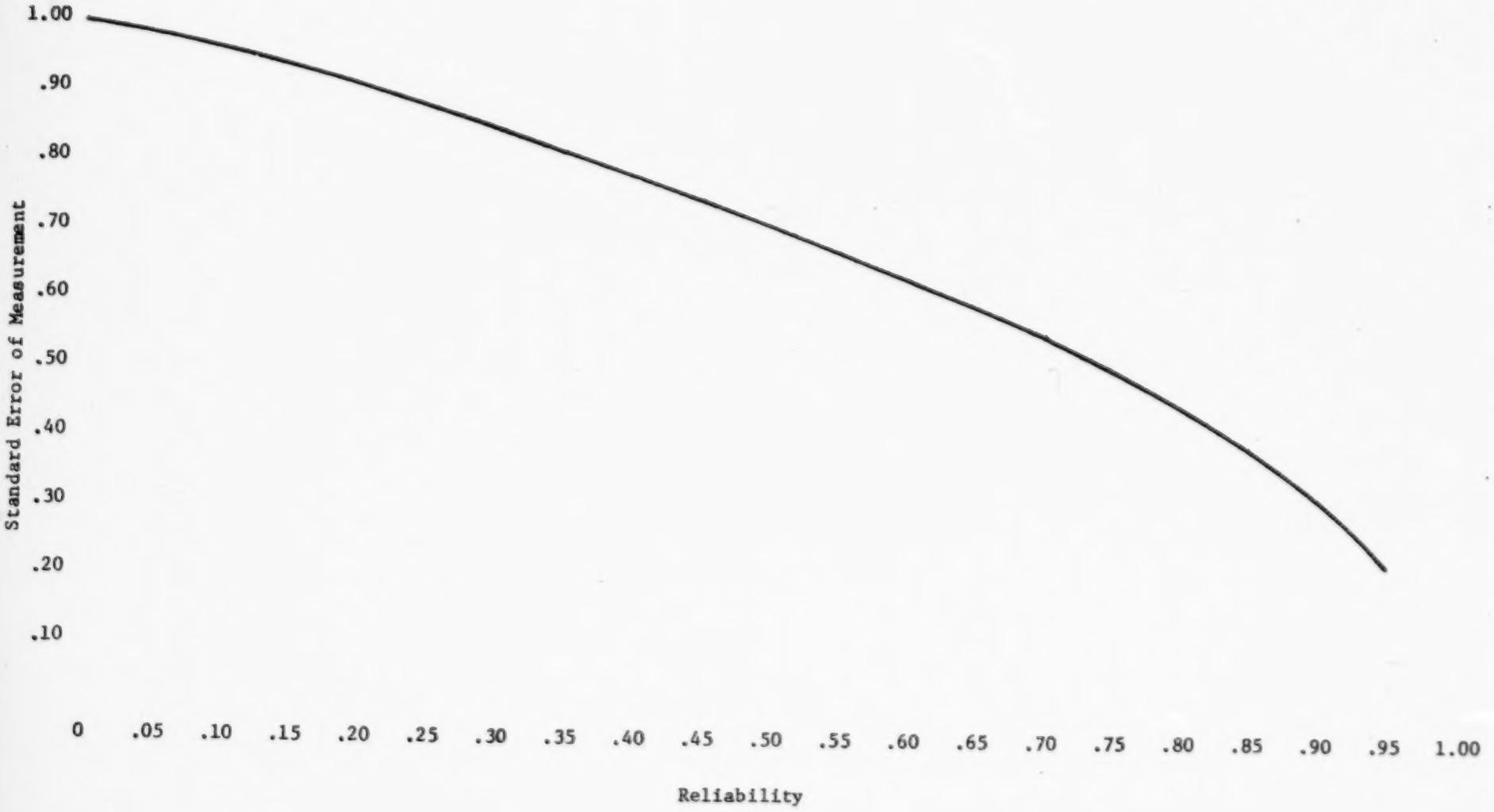
The magnitude of the S.E.M. for various size reliabilities is directly comparable to the decile cutoff scores. Thus, in Table 12, standardized cutoff scores for the deciles are presented. In Figure 2, the graph of standardized S.E.M.'s is presented as a function of reliability. For any reliability, the standardized S.E.M. can be determined. Then, one can assess for any standardized true score, how wide a plus or minus one S.E.M. interval is in deciles. He can also determine the interval for plus or minus two deciles, etc.

For example, from the graph, it can be seen that a reliability of .90 has a standardized S.E.M. of .31. Thus, an interval of plus or minus one S.E.M. is .62. An interval of plus or minus two S.E.M.'s is 1.24, etc. These values can be used in conjunction with Table 12. If an instructor's true score were in the center of the fifth decile, a standard score of .13, 68 percent of his observed scores would be from the standard score of -.19 (.13 - .32) to +.55 (.13 + .32), which is within the fourth decile to the sixth decile. Similarly, 95 percent of his observed scores would be from the third to the seventh decile.

In like manner, the reader can make his determinations for any size reliability and any true score. The reliability is a function, of course, of the item in which he is interested, and the number of raters.

In the case of subscales, the reliability can be an equivalence coefficient or a stability coefficient. The one which is used depends upon the question being asked. If one is concerned about how stable his ratings are likely to

Figure 2. Standardized Standard Error of Measurement as a Function of Reliability



be, i.e., to what extent they would be expected to vary from class to class, then he should determine his S.E.M. from stability coefficients. In the context of evaluation, this would seem to be the proper coefficient to use.

On the other hand, if the instructor is interested in how precisely a given attribute of teaching is measured, e.g., Interest and Attention, then equivalence coefficients are the proper coefficients for determination of the S.E.M. If the primary purpose is diagnosing the effect of a course in terms of various attributes, the equivalence coefficient would seem to be the proper coefficient.

NOTE: The results of this paper can also be used with the new CEQ Form 72 and Form 73 as it contains items and subscales found in Form 66.

References

Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297-334.

Ebel, R.L. Estimation of the reliability of ratings. *Psychometrika*, 1951, *16*, 407-424.

Eisenhart, C. Expression of uncertainties of final results. *Science*, 1968, *160*,1201-1204.

Gillmore, G.M. Approximating decile norms for the Illinois Course Evaluation Questionnaire by use of the normal curve. Research Report No. 342, Urbana, Illinois: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1972 (mimeo).

Gillmore, G.M. and Stallings W.M. A note on "accuracy" and "precision." *Journal of Educational Measurement*, 1971, *8*, 127-129.

Jensen, A.R. Note on why genetic correlations are not squared. *Psychological Bulletin*, 1971, *75*, 223-224.

Nunnally, J.C. *Tests and measurements: Assessment and prediction*. New York: McGraw-Hill, 1959.

Rozeboom, W.W. *Foundations of the theory of prediction*. Homewood, Illinois: The Dorsey Press, 1966.

Spencer, R.E. and Aleamoni. L.M. The Illinois Course Evaluation Questionnaire: A description of its development and a report of some of its results. Research Report No. 292. Urbana, Illinois: Measurement and Research Division, Office of Instructional Resources, 1969 (mimeo).

Treloar, A.E. *Correlation Analysis*. Minneapolis: Burgess Publishing Co., 1942.

Appendix A

The Illinois Course Evaluation Questionnaire (Form 66)

ILLINOIS COURSE EVALUATION QUESTIONNAIRE — FORM 66

Measurement and Research Division, Office of Instructional Resources, UNIVERSITY OF ILLINOIS © By Richard E. Spencer, 1965

IDENTIFICATION INFORMATION

STUDENT NUMBER

EXPECTED GRADE IN THIS COURSE

ARE YOU / IS THIS COURSE

TODAY'S DATE MO DAY YR

COURSE CODE

MARK YOUR COLLEGE: AGRIC. · COMM. & BA · EDUC. · ENGIN. · F & APP. ARTS · HOME EC · JOURN. COMM. · LAS · LAW · PHYS. ED. · MEDICINE · NURSING · PHARMACY · DENTISTRY · VET. MED. · OTHER

CAMPUS LOCATION: URBANA · CHI. CIR. · MED. CEN. · EXTENSION · OTHER

ARE YOU: FRESHMAN · SOPHOMORE · JUNIOR · SENIOR · GRAD. · OTHER

DIRECTIONS:
1. PRINT INSTRUCTOR'S LAST NAME HERE_____
2. COMPLETE INDENTIFICATION INFORMATION TO THE RIGHT.
3. RESPOND TO THE ITEMS PRESENTED FRANKLY AND COMPLETELY-ONE RESPONSE PER ITEM. (SEE SAMPLE MARK AND RESPONSE CODE)
4. USE PENCIL ONLY-DO NOT USE PEN, BALL POINT OR INK OF ANY KIND.

SAMPLE MARKS:

USE PENCIL ONLY

a) ▮
b) ▮
c) ▮
d) ▮

RESPONSE CODE:

MARK SA — IF YOU STRONGLY AGREE WITH THE ITEM
MARK A — IF YOU AGREE MODERATELY WITH THE ITEM
MARK D — IF YOU DISAGREE MODERATELY WITH THE ITEM
MARK SD — IF YOU STRONGLY DISAGREE WITH THE ITEM

IF PART II OR III IS TO BE USED MARK HERE →

COMPLETE SECTIONS BELOW ACCORDING TO YOUR INSTRUCTOR'S DIRECTIONS:

OPTIONAL PART II ITEMS 51-75

OPTIONAL PART III ITEMS 76-100

1. I learn more when other teaching methods are used.
2. It was a waste of time.
3. Overall, the course was good.
4. The textbook was very good.
5. The instructor seemed to be interested in students as persons.
6. More courses should be taught this way.
7. The course held my interest.
8. I would have preferred another method of teaching in this course.
9. It was easy to remain attentive.
10. The instructor did not synthesize, integrate or summarize effectively.
11. Not much was gained by taking this course.
12. The instructor encouraged the development of new viewpoints and appreciations.
13. The course material seemed worthwhile.
14. It was difficult to remain attentive.
15. Instructor did not review promptly and in such a way that students could understand their weaknesses.
16. Homework assignments were helpful in understanding the course.
17. There was not enough student participation for this type of course.
18. The instructor had a thorough knowledge of his subject matter.
19. The content of the course was good.
20. The course increased my general knowledge.
21. The types of test questions used were good.
22. Held my attention throughout the course.
23. The demands of the students were not considered by the instructor.
24. Uninteresting course.
25. It was a very worthwhile course.
26. Some things were not explained very well.
27. The way in which this course was taught results in better student learning.
28. The course material was too difficult.
29. One of my poorest courses.
30. Material in the course was easy to follow.
31. The instructor seemed to consider teaching as a chore or routine activity.
32. More outside reading is necessary.
33. Course material was poorly organized.
34. Course was not very helpful.
35. It was quite interesting.
36. I think that the course was taught quite well.
37. I would prefer a different method of instruction.
38. The pace of the course was too slow.
39. At times I was confused.
40. Excellent course content.
41. The examinations were too difficult.
42. Generally, the course was well organized.
43. Ideas and concepts were developed too rapidly.
44. The content of the course was too elementary.
45. Some days I was not very interested in this course.
46. It was quite boring.
47. The instructor exhibited professional dignity and bearing in the classroom.
48. Another method of instruction should have been employed.
49. The course was quite useful.
50. I would take another course that was taught this way.

"PERMISSION TO REPRODUCE THIS COPY RIGHTED MATERIAL HAS BEEN GRANTED BY Illinois University TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER"

OPTICAL SCANNING FORM by OPTICAL SCANNING CORPORATION © 1965